Learning Statistics Secure from Administrative to DNA Records: Are We There Yet?

Bradley Malin, Ph.D.
Professor of Biomedical Informatics, Biostatistics, & Computer Science
Director, Health Data Science Center
Vanderbilt University
February 24, 2017

What are We Talking About?

- Getting *answers* out of data…

  … without revealing individual records

- Not a new concept (remember the ‘70s?)!

- But now the computational machinery may be powerful enough for prime time

Department of Education

First	Last	Age	Education
John	Smith	32	High School
Jim	Jones	45	College
Mary	Little	39	Medical School
Mike	Glasgow	22	College
Abby	Hightower	51	Medical School
Raj	Ramesh	62	College

Internal Revenue Service

First	Last	Age	Earnings
Tyler	Tooney	27	$45,000
Jim	Jones	45	$60,000
Mary	Little	39	$150,000
Bill	Blast	75	$275,000
Abby	Hightower	51	$75,000
Sandy	Tunep	62	$66,000

Traditional Cryptography

Department of Education

6	12	19	23

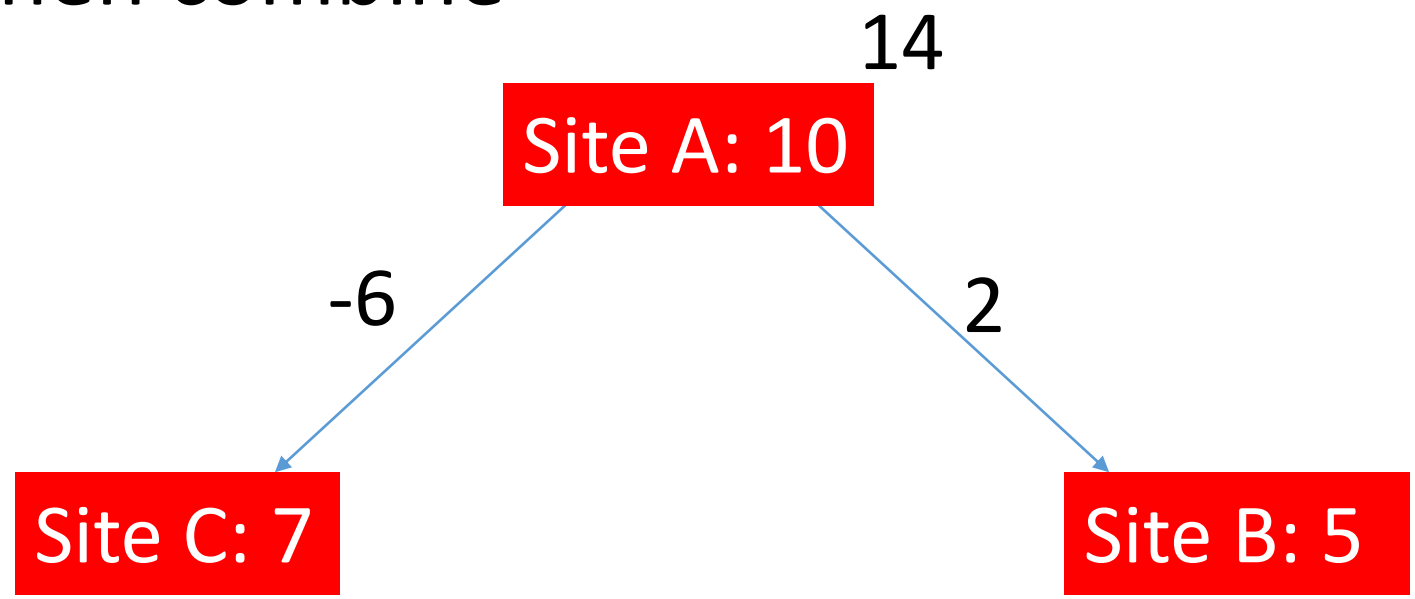
Internal Revenue Service

6	30	33	39

This *leaks* alot of information

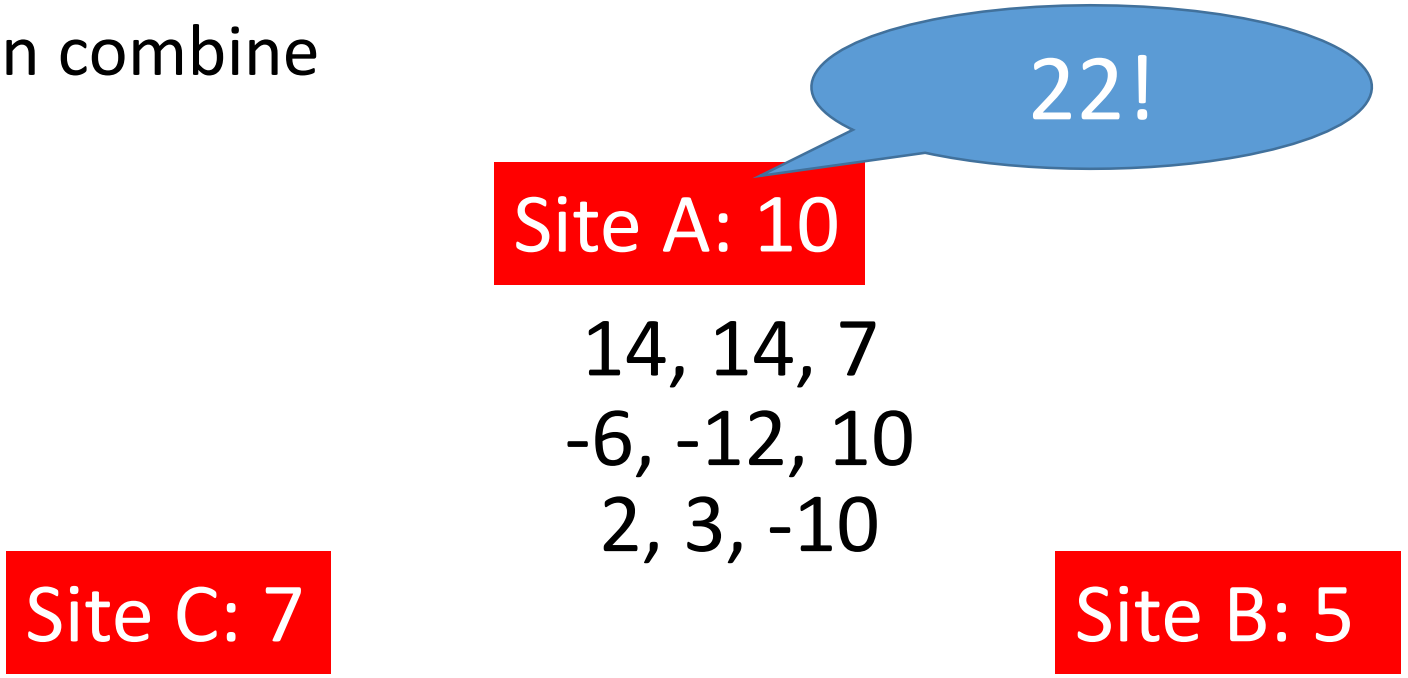
Crypto Advances: Secret Sharing

- Split… then combine



Crypto Advances: Secret Sharing

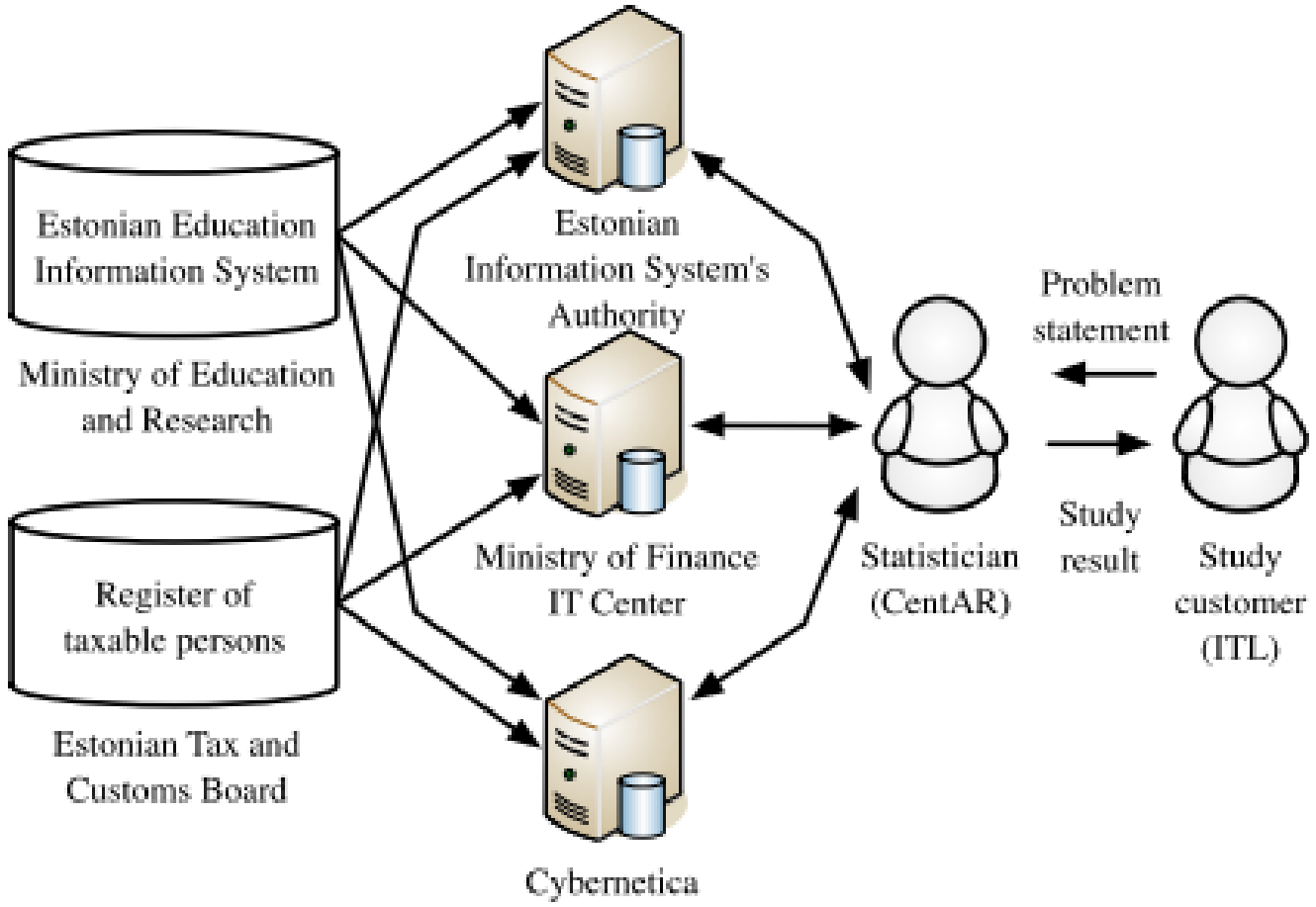
- Split… then combine



- In practice, this is done through higher order functions
- And once you can count – you can create complex statistical models

Real World Example: Cybernetica

- Statisticians from the Estonian Center of Applied Research
- Sharemind System
- Linked:
  - Individual tax payments from Estonian Tax and Customs Board (10 million)
  - Higher education events from Ministry of Education and Research (500 thousand)



D. Bogdonov, et al. Students and taxes: a privacy-preserving social study using secure computation. Proceedings of the Privacy Enhancing Technologies Symposium. 2016.

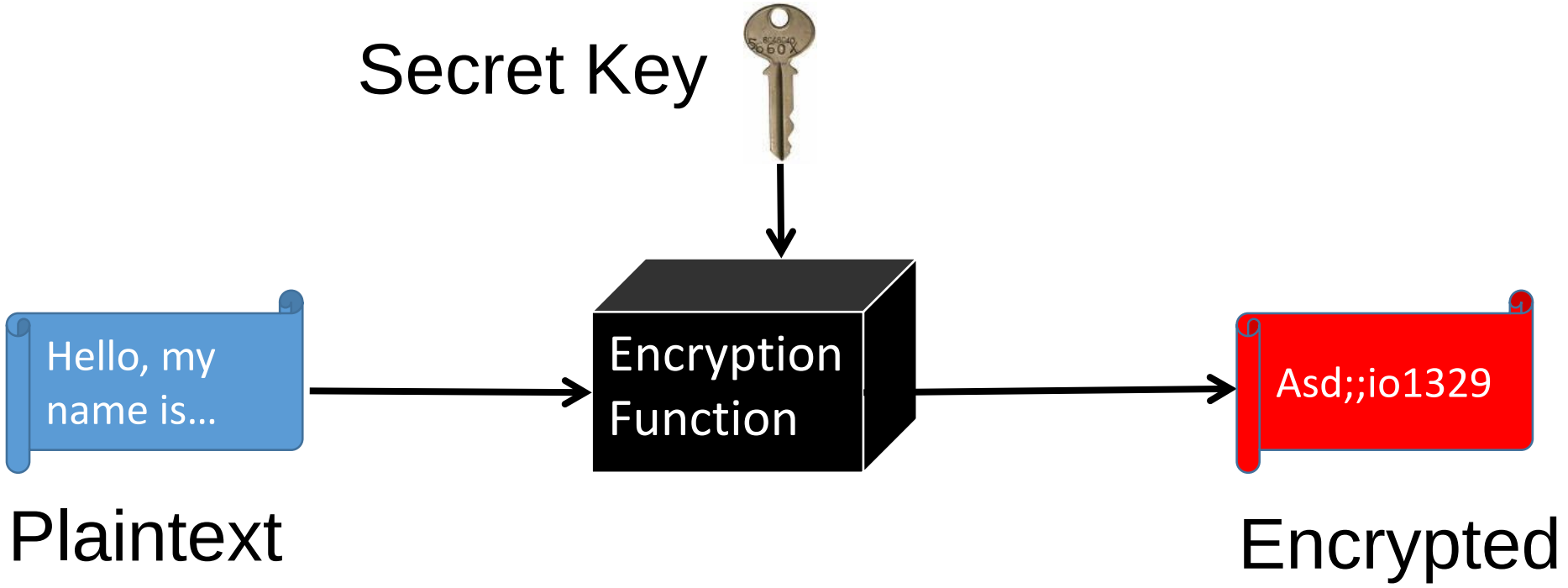
Real World Example: Cybernetica (Caveat)

Process	Time (testing facility)	Time (in the wild)
Aggregation of education data	30 minutes	2 hours
Aggregation of tax data (monthly income)	18 hours	221 hours
Aggregation of tax data (yearly income)	2 hours	15 hours
Data join	30 minutes	4 hours
Analysis of data	29 hours	141 hours
Total time	**60 hours**	**384 hours**

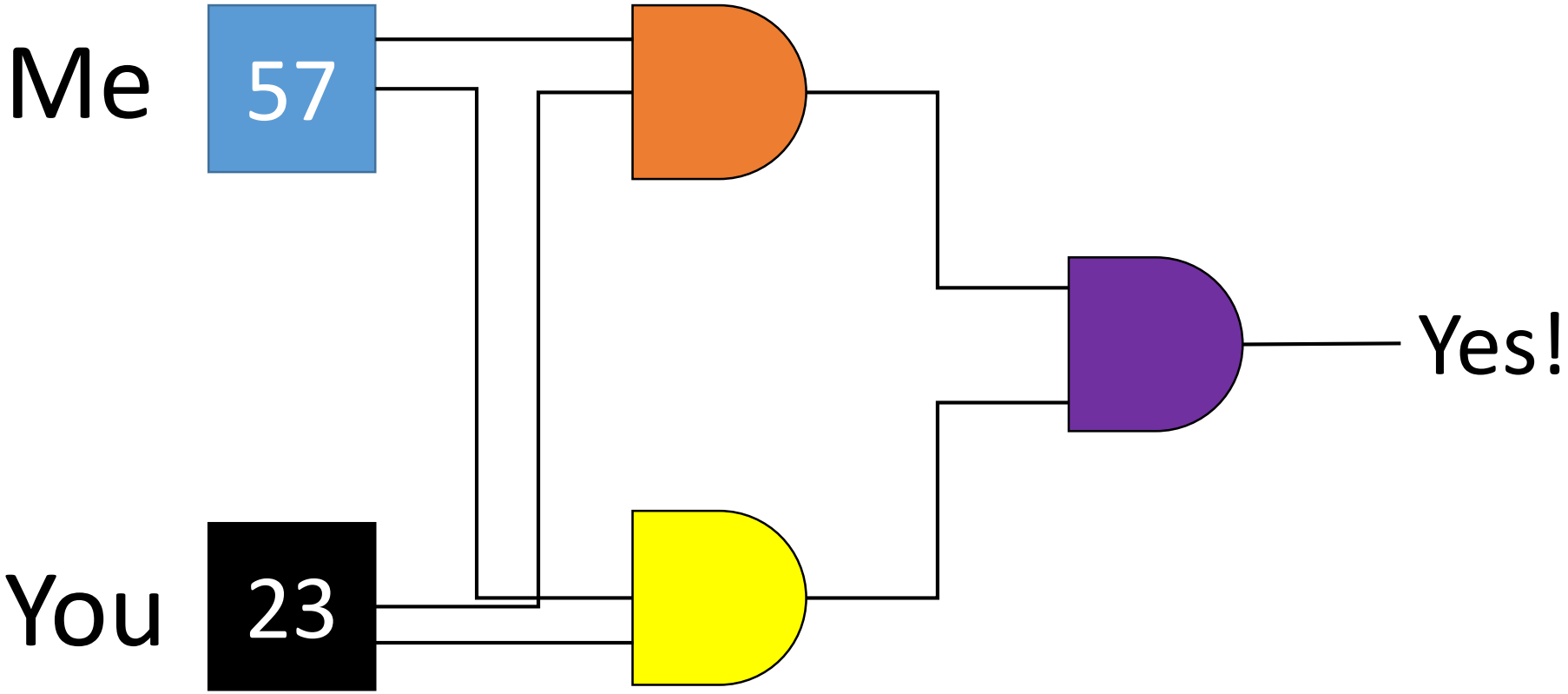
D. Bogdonov, et al. Students and taxes: a privacy-preserving social study using secure computation. Proceedings of the Privacy Enhancing Technologies Symposium. 2016.

What Happens When You Can’t Find Trusted Servers?

Traditional Cryptography



Crypto Advance - Circuit Evaluation:
Am I Older than You?



“Garbled” Circuit Evaluation

❶ Encrypted input

❷ Process encrypted data

❸ Extend circuit and carefully randomize order

“Garbled” Circuit Evaluation

The circuits can get big...

Process encrypted data

The circuits can get slow.

Crypto Advances: Homomorphisms

- Long word, simple idea

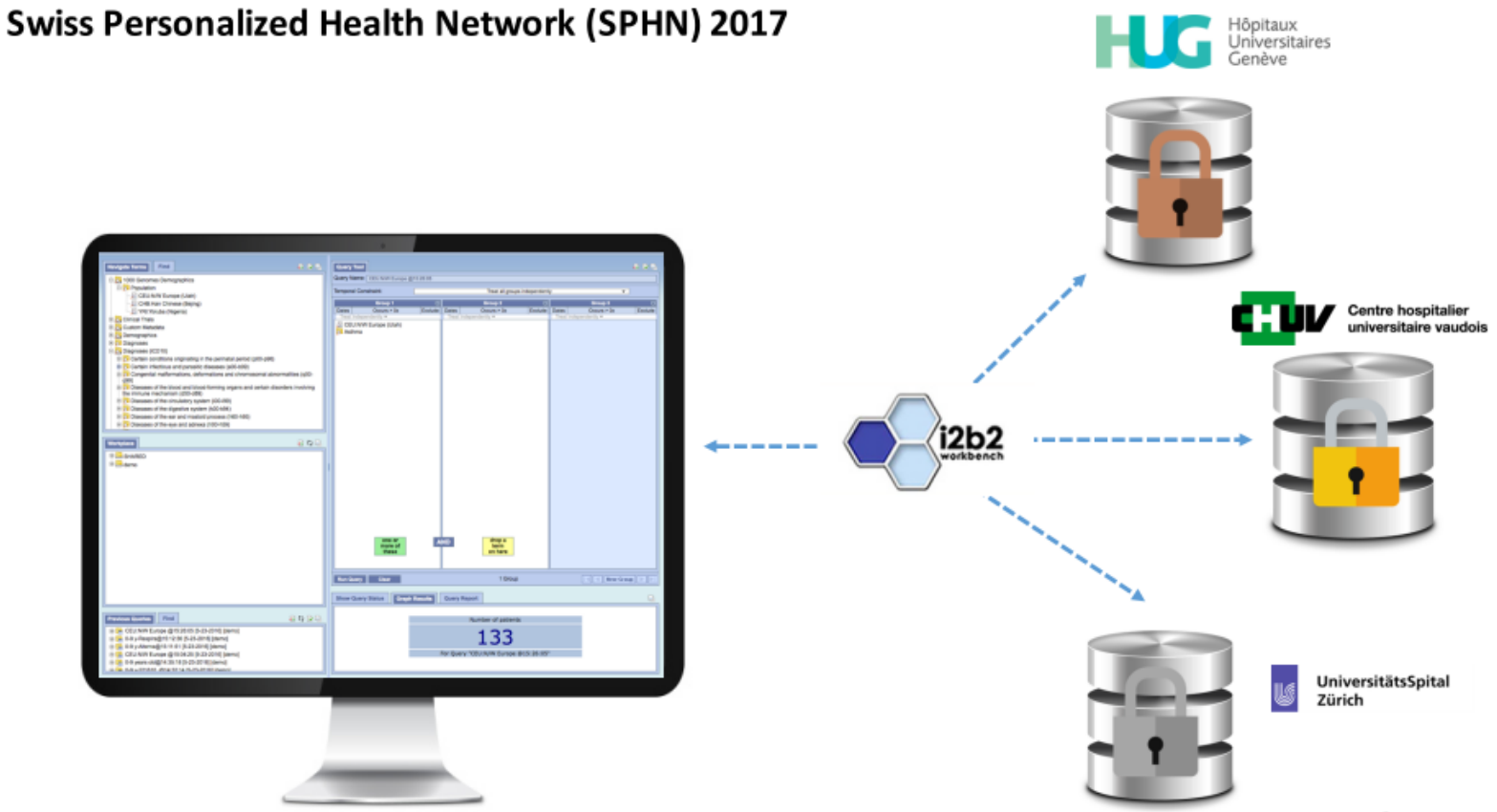
$$\text{Encrypt}(X + Y) = \text{Encrypt}(X) + \text{Encrypt}(Y)$$

$$\text{Decrypt}(\text{Encrypt}(X + Y) = X + Y$$

- Can perform arbitrary mathematical computations!
- Major recent advances in academia and industry.

Application in Health Data Mining

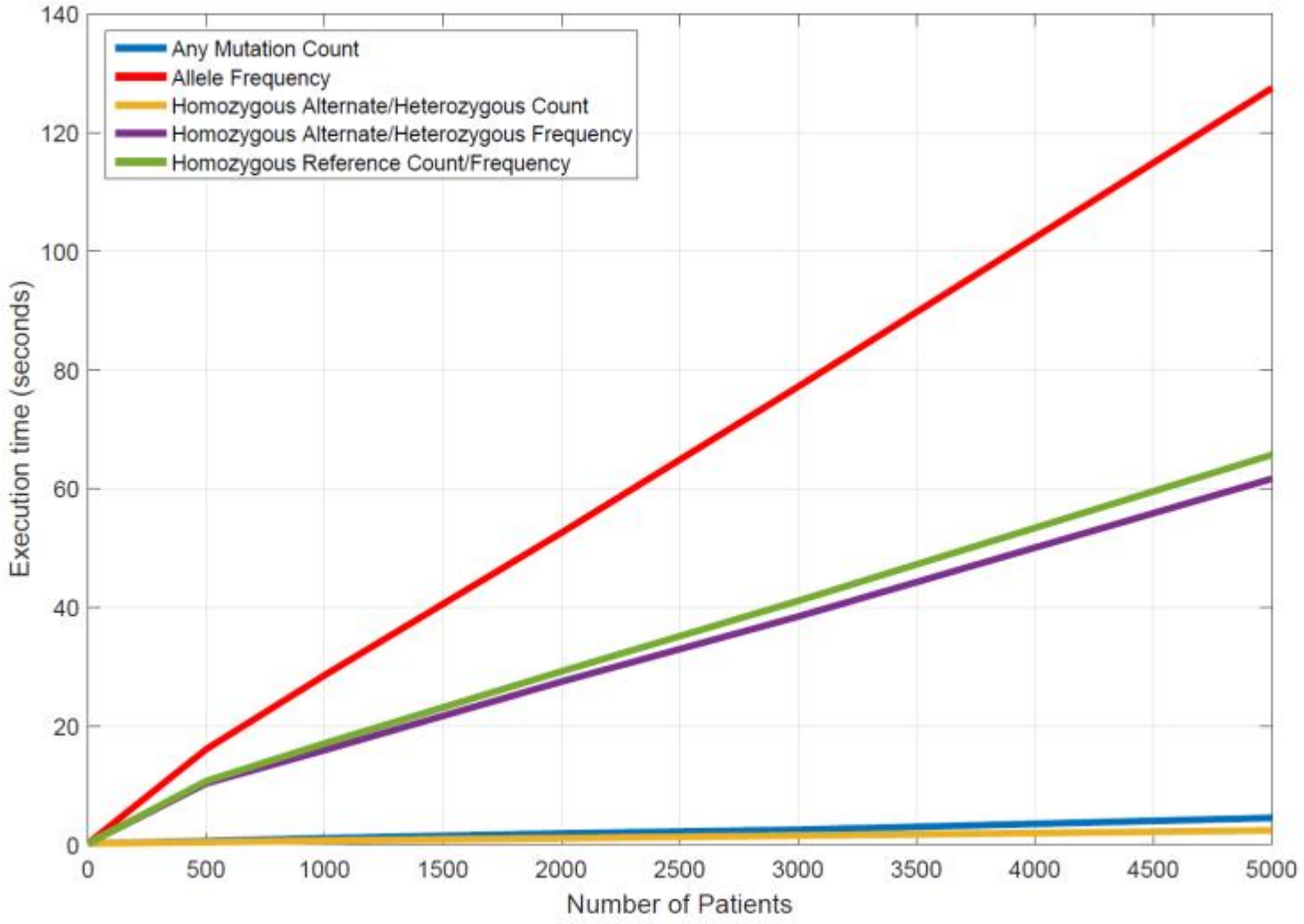
- EPFL & Sophia Genetics
- Integration of homomorphic crypto into popular health database exploration software (i2b2)
- Issue queries like “how many patients have congestive heart failure and genetic variant x?”



J.L. Raisaro, et al. Privacy-preserving explocation of genetic cohorts with i2b2 at Lausanne University Hospital. 3rd International Workshop on Genome Privacy & Security. 2016.

Application in Health Data Mining

- EPFL & Sophia Genetics
- Integration of homomorphic crypto into popular health database exploration software (i2b2)
- Issue queries like “how many patients have congestive heart failure and genetic variant x?”



J.L. Raisaro, et al. Privacy-preserving explocation of genetic cohorts with i2b2 at Lausanne University Hospital. 3rd International Workshop on Genome Privacy & Security. 2016.

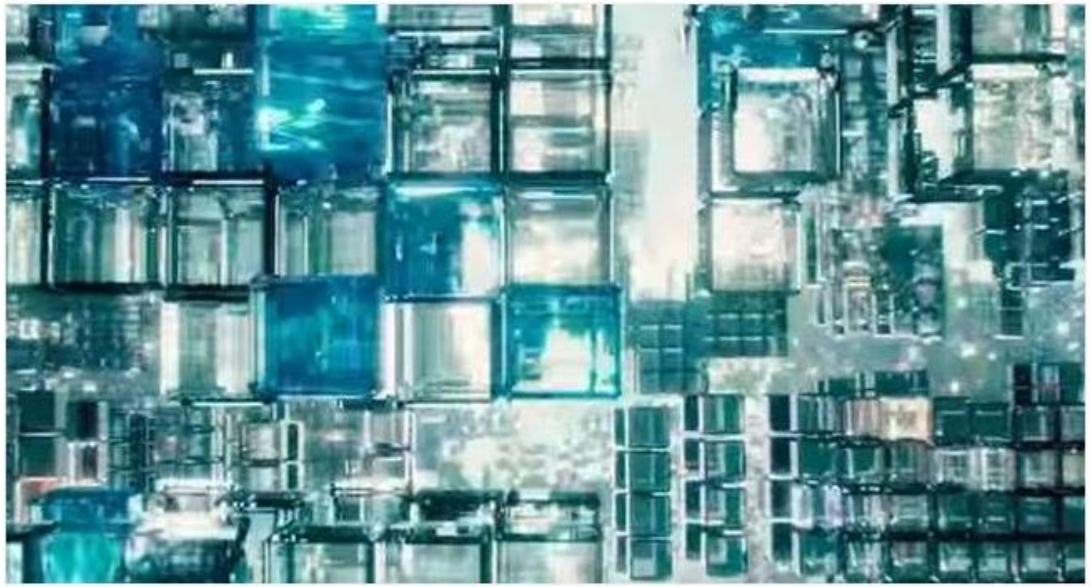
The Register®
Biting the hand that feeds IT

DATA CENTER SOFTWARE SECURITY TRANSFORMATION DEVOPS BUSINESS PERSONAL TECH

Software

Microsoft researchers smash homomorphic encryption speed barrier

Artificial intelligence CryptoNets chew data fast but keep it safe



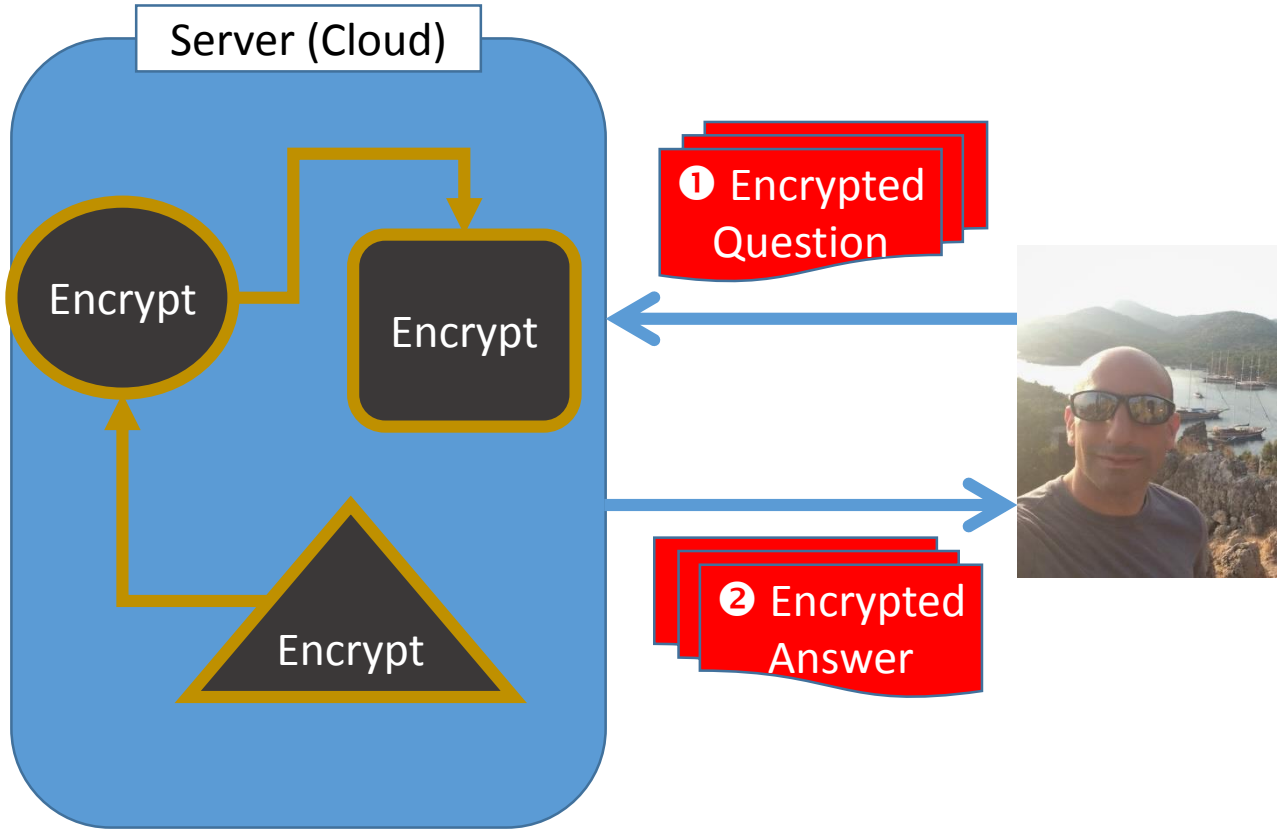
Ultron it isn't, thank goodness

9 Feb 2016 at 08:02, Iain Thomson

Exclusive Microsoft researchers, in partnership with academia, have published a paper detailing how they have dramatically increased the speed of homomorphic encryption systems.

With a standard encryption system, data is scrambled and then decrypted when it needs to be processed, leaving it vulnerable to theft. Homomorphic encryption, first proposed in 1978 but only really refined in the last decade thanks to increasing computing power, allows

https://www.theregister.co.uk/2016/02/09/researchers_break_homomorphic_encryption/



Real Working Software

- Time to compute is quick – seconds (or less)!

- I encourage you to watch Kristen Lauter's talk from the 2016 Genomics and Patient Privacy Conference

https://www.youtube.com/watch?v=vUtyuw7YLVM

The following video is presentation from Kristin Lauter, PhD, Principal Researcher and Research Manager for the Cryptography group at Microsoft Research. She presents a demonstration of privacy preserving technologies for genomic data sharing.

Kristin's group has won the iDASH 2016 competition in Track 3, Testing for Genetic Diseases on Encrypted Genomes (secure outsourcing). This is to calculate the

Where are We?

- The basic math behind secure computation is there.
- The software is there … for special circumstances.
- The software is not there… for arbitrary on the fly computations*

**It is for secret sharing, but not for homomorphic computation*

What are the Challenges?

- Secure computation is NOT a panacea
- Computation can be secured, but the answers can still leak information
  - Example: Queries that reveal answers with very small counts
- There will always be a need for good key management, authentication, and (to a certain extent) trust – both in the system and the data

One More Thing – Secure Hardware

- Multiparty crypto may not be necessary in the future.
- Tamper resistant hardware may provide an opportunity for performing statistical analysis on plaintext.
- Hardware obscures individuals from viewing what's taking place.
- Example: our work on the IBM Secure co-processor in 2012*

*Canim M, Kantarcioglu K, Malin B. Secure management of biomedical data with cryptographic hardware. IEEE Transactions on Information technology in biomedicine. 2012; 16(1): 166-175.

- More recently: UCSD's work on the Intel SGX system (secured RAM)**

**F. Chen, et al. PRINCESS: Privacy-protecting rare disease international collaboration via encryption through software guard extensions. Bioinformatics. 2017: in press